研究生数学教学系列（农林类） 10

多元统计分析

袁志发　周静芋　主编

GM

科学出版社

研究生数学教学系列(农林类)

多元统计分析

袁志发　周静芋　主编

科学出版社

北　京

内 容 简 介

多元统计分析是从经典统计学中发展起来的一个分支，是一种综合分析方法，它能够在多个对象和多个指标互相关联的情况下分析它们的统计规律，很适合农业科学研究的特点．本书是高等农、林院校研究生多元统计分析教材．主要内容包括多元正态分布及其抽样分布、多元正态总体的均值向量和协方差阵的假设检验、多元方差分析、直线回归与相关、多元线性回归与相关(Ⅰ)和(Ⅱ)、主成分分析与因子分析、判别分析与聚类分析、Shannon信息量及其应用．

本书适合作农、林院校研究生教材，亦可供高等院校高年级学生及教师和农林科技工作者参考．

图书在版编目(CIP)数据

多元统计分析/袁志发，周静芋主编．—北京：科学出版社，2002
研究生数学教学系列(农林类)
ISBN 7-03-010798-5

Ⅰ．多… Ⅱ．①袁…②周… Ⅲ．多元分析：统计分析 Ⅳ．O212．4

中国版本图书馆CIP数据核字(2002)第073487号

责任编辑：李 锋 陈玉琢/责任校对：钟 洋
责任印制：安春生/封面设计：黄华斌 王 浩

科学出版社出版
北京东黄城根北街16号
邮政编码：100717
http://www.sciencep.com
新蕾印刷厂 印刷
科学出版社总发行 各地新华书店经销

2002年10月第 一 版 开本：B5(720×1000)
2002年10月第一次印刷 印张：19 1/2
印数：1—3 500 字数：382 000

定价：28.00元

(如有印装质量问题，我社负责调换〈环伟〉)

前　言

作者从20世纪80年代末便从事有关生物数学、数量遗传学和群体遗传学等方面的研究生教学工作，后来又承担了动物遗传育种与繁殖专业的硕士生、博士生及应用数学硕士生的培养工作.在长期的科研和教学中，作者思考着农业科学的发展与数学发展的关系问题，因为这个问题直接关系着解决农、林院校研究生数学教学的内容和体系问题.

农业科学是现代科学技术中应用最广阔、最活跃、最富挑战性的领域之一.追根溯源，它与数学的发展，尤其与统计学的发展，具有同步性.数学与农业、管理的关系是从人类计数活动开始的，正如管仲所说"不明于计数，犹如无舟楫欲径于水，险也".近代农业应用数学是由于生物、工程和经济等科学的进步而发展的.随着19世纪初近代生物学和经济统计学的进步，导致了20世纪初遗传学、经济学与数学的融合和交叉.生物学家认为"生物学(biology)这个名词来源于希腊字 oiooto (生命)，这门科学由于应用了数学，获得了第二次生命".列宁认为"统计学家和经济学家各走各的路，那么他们两者都不能获得满意的结果".20世纪初学科间的融合和以后的发展，使农业应用数学形成了与生物数学、经济数学、工业数学相平行而互相交融的发展局面.从与农业科学交叉的意义上看，有数量分类学、群体遗传学、数量遗传学、数量生态学、数量生理学、数量经济学、生物信息学、农业系统工程学等.从数学方法上讲，有统计学、信息论、系统论、控制论、生物方程、运筹学等.概括起来，农业应用数学是农业领域中可应用的数学.从含义上讲，有三个方面：一是应用数学知识来解决农业中的实际问题，以求实效，它包括为此而建立的数学模型、计算机模拟法研制等；二是与农业科学相交叉，形成新的学科.如群体遗传学、生物信息学等；三是从农业科学中提炼出数学问题进行研究，从而发展数学理论.如基因如何从时间、空间上来精细地控制发育过程等.

从农业科学的研究特点上看，它是以实验和调查为前提的研究过程.首先是根据研究目的进行周密而审慎的试验设计或抽样设计，通过实施而得到数据，如孟德尔的豌豆实验、摩尔根的果蝇实验、田间调查等.然后通过试验设计或抽样设计的数学模型，进行分析而得到研究结论，其中包括了刻画指标之间关系的数学模型研究.在研究过程中，数学方法起到了把实验数据转化为研究结论的作用，如试验设计的数学模型起到了把处理转化为输出指标的作用.又如把样品的指标观察值转化为分类的结果等.

经过20余年的探索，并不断地总结经验和教训，根据农业科学发展的历史和

现代农业科学的需要，高等农、林院校开设研究生数学课程体系和内容已初步形成.在统计学方面，开设数理统计、试验设计与分析、多元统计等；在生长分析方面，讲授生长方程、室分析等动力学模型；其它还有模糊数学方法、信息论方法和运筹学等有关内容.生物信息论或计算生物学将成为研究生的一门新课程.

值得强调的是，自 1978 年以来，农、林院校的同仁们在农业应用数学的研究上取得了很大进展，并且出版了相当数量的著作.在试验设计方面，有赵仁容、马育华、俞渭江、莫惠栋、徐中儒和袁志发与周静芋等人的著作；在数理统计上有符伍儒、袁志发与顾天骥和朱军等人的著作；在多元统计分析方面，有裴鑫德、袁志发与孟德顺等人的著作；在数量遗传方面，有吴仲贤、马育华、吴常信、盛志廉、刘来福、裴新澍、高之仁、兰斌与袁志发、朱军、胡秉民和徐碧云等人的著作与译著；在模糊数学方面，有袁志发、杨崇瑞等人的著作.另外，1985 年《生物数学学报》的创刊、1987 年农学会农业应用数学分会的成立及农业应用数学硕士点的建立(如西北农林科技大学)等，使农业应用数学的发展进入了一个新的阶段.

多元统计分析是从经典统计学中发展起来的一个分支.多元统计分析是一种综合分析方法，它能够在多个研究对象和多个指标互相关联的情况下分析出它们的统计规律，很适合农业科学研究的特点.另外，多样性问题是农业科学中很重要的研究内容，如生物多样性，物质运动和形式的多样性等.多样性的度量可以是统计学的，亦可以是信息论的，但从本质上讲应该由信息方法来度量.因此，在本教材中，加入了 Shannon 信息量及其应用一章.这样做的好处之一是统计分析和信息分析可以互为借鉴，如统计学中的关联分析(回归、相关等)与信息论中的互信息和离散增量分析等；另外，还可以与分子生物学分析相呼应，开阔读者的视野.

本书由西北农林科技大学、南京农业大学和安徽农业大学联合完成.主编为袁志发、周静芋；副主编有周宏、卢恩双、吴坚、宋世德、郭满才；参加编写的有汪小龙、孙世铎、雷雪芹、刘建军、张宏礼、解小莉、李相运、郑会玲、梅拥军；郑瑶绘图.

编写一本好的研究生教材是很难的.尽管作者是学习数学的，并且从事农业科学研究 30 余年，仍感力不从心.本书难免有遗漏和不妥之处，望同仁和读者批评指正.

袁志发　周静芋

2000 年 4 月于杨凌

目　录

第一章　多元正态分布及其抽样分布

§1.1　多元指标统计数据及其图示

1.1.1　多元统计数据

统计学数据是通过实验或调查而得到的.统计学中把一个随机变量称为一维总体,把多维随机变量称为多维总体,相应的变量称为一维总体变量和多维总体变量.总体是由若干个元素组成的集合,每个元素称为个体,每个个体的数量或非数量(质量)特性由总体变量来刻画.总体中所含个数的数目称为总体容量,用 N 表示,当 N 有限时,称为有限总体,否则称为无限总体.统计学中,习惯上用随机变量 X, Y 等表示总体,为了对总体 X 的分布规律及其特性进行研究,最好的办法是把总体上所有的个体都在同一条件下测定,然后进行分析,这往往是难以实现的,因为对于无限总体办不到,对有限总体的每个个体进行破坏性测定是不允许的.可行的办法是对总体进行抽样观测,对总体的分布和特性进行估计,从总体中随机抽取 n 个个体,其总体变量分别为 $X_1, X_2, \cdots, X_n$,称为总体 X 的容量为 n 的简单随机样本."随机抽取"是指总体中每个个体都有相同的机会进入样本,这样的样本才能客观地反映总体,同时保证了 $X_i(i=1,2,\cdots,n)$ 与 X 同分布且 X_i 间相互独立.当样本被实际测定时,所得的是 n 个实际的数据 $x_1, x_2, \cdots, x_n$,它称为样本点或样本值.一般来讲,同一样本的不同次的实际抽取观察测定得到的样本值是不同的.样本的一次实际抽取测定称为样本在一次观测中的实现.显然,一个样本可有无限次的观测实现.用样本资料推断总体规律的方法,是数理统计分析的任务,亦是统计方法的特点.

总体变量可分为三种类型:

1.名称属性(nominal attribute)

名称属性是用名称把总体中各个个体描述为若干不同的状态,每个个体具有一种状态,各状态之间无一定顺序.如土壤的颜色可分为红、黑、黄等;又如植被可分为森林、草原、灌丛、苔原等.

名称属性按其状态多少又分为两类:

①二元属性(binary attribute)

二元名称属性只有对立的两种状态,如昆虫有翼无翼,某植物有刺无刺等.

②无序多状态属性(disordered multistate attributes)

这种属性系指具有三个以上无序状态的名称属性.具有 n 个状态的无序多状

态属性可分解为 n 个二元属性. 如土壤颜色具有红、黑、黄三个状态, 可转化为三个二元属性(表 1.1).

表 1.1　土壤颜色的分解量化

状态 / 多元属性	红	黑	黄
1(红与非红)	1	0	0
2(黑与非黑)	0	1	0
3(黄与非黄)	0	0	1

2. 顺序属性(ordinal attribute)

具有多种顺序状态的属性称为顺序属性. 如土壤酸碱度分为强酸性、弱酸性、中性、弱碱性和强碱性五个状态, 可用 1, 2, 3, 4, 5 表示各状态. 又如植物种子分为大、中、小三级, 可用 1, 2, 3 来表示. 由于顺序间的差距没有明确表示, 故用 1, 2, 3, …参加运算亦不适宜. 如果顺序属性是用数量来划分的, 最好用数量来表示为宜. 此类数据的转化是较为麻烦的.

3. 数量属性(guantitative attribute)

用数值来表示的属性称为数量属性. 如重量、长度等.

数量属性可分为离散数量属性和连续数量属性两类. 如一簇上开几朵花为离散数据, 只能取 0, 1, 2, …整数值. 而重量、深度为连续性数量性状.

有的将数量属性分为比例量和区间量两种, 像长度、轻重这类量, 0 是有明显的物理意义的, 可以说 10 克是 5 克的 2 倍, 故说它们是比例量. 另外像温度、时间等, 0 仅为计数标准, 这类量仅有等间隔性质, 无比例性质, 10℃ 与 20℃ 之间与 0℃ 与 10℃ 之差为间隔, 但不能说 20℃ 比 10℃ 热一倍, 故它们只能是区间变量.

为了便于分析, 要求将调查或测定的一组个体的若干个属性的原始数据, 排列成规定的形式. 一般情况下, 我们假定有 n 个个体, p 个属性的样本排成如表 1.2 形式:

表 1.2　抽样数据

个体 / 属性	$X_{(1)}$	$X_{(2)}$	…	$X_{(n)}$
1	X_{11}	X_{12}	…	X_{1n}
2	X_{21}	X_{22}	…	X_{2n}
⋮	⋮	⋮		
p	X_{p1}	X_{p2}	…	X_{pn}

上述样本可表示成矩阵形式

$$\underset{p\times n}{X}=(X_{(1)},X_{(2)},\cdots,X_{(n)})=\begin{bmatrix}X_{11} & X_{12} & \cdots & X_{1n}\\ X_{21} & X_{22} & \cdots & X_{2n}\\ \vdots & \vdots & & \vdots\\ X_{p1} & X_{p2} & \cdots & X_{pn}\end{bmatrix}$$

其中 X_{ij} 为样本中第 j 个个体的总体变量 $X_{(j)}$ 的第 i 个分变量. 若为样本值，将 X_{ij} 换成 x_{ij}. 显然，$X_{(j)}$ 是第 j 个个体的 p 维随机向量(样本)或 p 维统计数据向量(实现值).

【例 1】 桔梗科六个种的性状原始数据如表 1.3 所示.

表 1.3 桔梗科六个种的抽样数据

属性		茎是否缠绕	株高(m)	叶序	叶缘	花序	子房室数	果裂方式	种子是否具翼
名称	样品号	1	2	3	4	5	6	7	8
党参	1	1	5.5	1	0	0	4	2	0
桔梗	2	0	0.6	0	1	0	5	1	0
轮叶沙参	3	0	0.5	2	1	2	3	0	0
荠苨	4	0	0.7	0	2	1	3	0	0
羊乳	5	1	2.5	1	0	0	4	2	1
石沙参	6	0	0.65	0	1	2	3	0	0

其矩阵形式为

$$X=\begin{bmatrix}1 & 5.5 & 1 & 0 & 0 & 4 & 2 & 0\\ 0 & 0.6 & 0 & 1 & 0 & 5 & 1 & 0\\ 0 & 0.5 & 2 & 1 & 2 & 3 & 0 & 0\\ 0 & 0.7 & 0 & 2 & 1 & 3 & 0 & 0\\ 1 & 2.5 & 1 & 0 & 0 & 4 & 2 & 1\\ 0 & 0.65 & 0 & 1 & 2 & 3 & 0 & 0\end{bmatrix}$$

【例 2】 两种肉鸡在 16 周龄时的脂肪细胞(50 个细胞)、胸肌纤维(40 根)和腿肌纤维(40 根)的直径平均数据见表 1.4.

表 1.4 两个鸡种的脂肪和肌纤维直径数据(单位：μm)

性状 / 品种	脂肪细胞直径	胸肌纤维直径	腿肌纤维直径
艾维茵	60.50	52.48	53.10
星杂 882	74.55	42.48	51.94

1.1.2　多元数据的图示

图形是直观而形象的，它可以帮助人们思维和判断. 当只有两个变量时，通常用直角坐标在平面上点图；当有三个变量时，虽然可以在三维坐标里点图，但很不方便；当变量多于三个时，用通常的方法已不能点图了. 多元数据的图表示在20世纪70年代有了突破，许多方法应运而生，这里介绍两种简单实用的方法.

1. 雷达图

雷达图也称星图或蜘蛛网图，其作图步骤如下：划一个圆，当数据为 p 个时间的数据或一个时间断面上的 p 维数据时，把圆周用 p 个点等分，由圆心连接 p 个点，得 p 个辐射状的半径，将 p 个半径看做 p 个坐标轴，各坐标轴各标一个性状或特性，每个坐标轴的刻度可按各变量单位大小或标准化刻度而定. 将每个样品的变量值点刻在各自的坐标轴上，依次连接成一个 p 边形，这样就得到了这个样品的雷达图.

例1的雷达图如图1-1所示，具体作法如下：

1) 对 p 个属性的每一个进行标准化，把每个属性取值变到[0,1]区间内，最简单的标准化方法是极差标准化法. 设第 i 属性的数据为

$$x_{i1}, x_{i2}, \cdots, x_{in} \qquad (i = 1, 2, \cdots, p)$$

则极差标准化了的数据为

$$X'_{ij} = \frac{x_{ij} - \min\{x_{ij}\}}{\max\{x_{ij}\} - \min\{x_{ij}\}} \qquad (i = 1, 2, \cdots, n) \tag{1.1.1}$$

例1数据经极差标准化列在表1.5中.

表1.5　例1数据的极差标准化结果

属性		茎是否缠绕	株高/m	叶序	叶缘	花序	子房室数	果裂方式	种子是否具翼
名称	样品号	1	2	3	4	5	6	7	8
党参	1	1	1	0.5	0	0	0.5	1	0
桔梗	2	0	0.02	0	0.5	0	1	0.5	0
轮叶沙参	3	0	0	1	0.5	1	0	0	0
荠苨	4	0	0.04	0	1	0.5	0	0	0
羊乳	5	1	0.40	0.5	0	0	0.5	1	1
石沙参	6	0	0.03	0	0.5	1	0	0	0

2) 作六个单位圆且八等分，作雷达图1-1.

例2数据的雷达图如图1-2所示. 它是用原始数据作图的. 首先将圆周三等分，每个分点标所示属性的最大值，然后按数据比例作图而成.

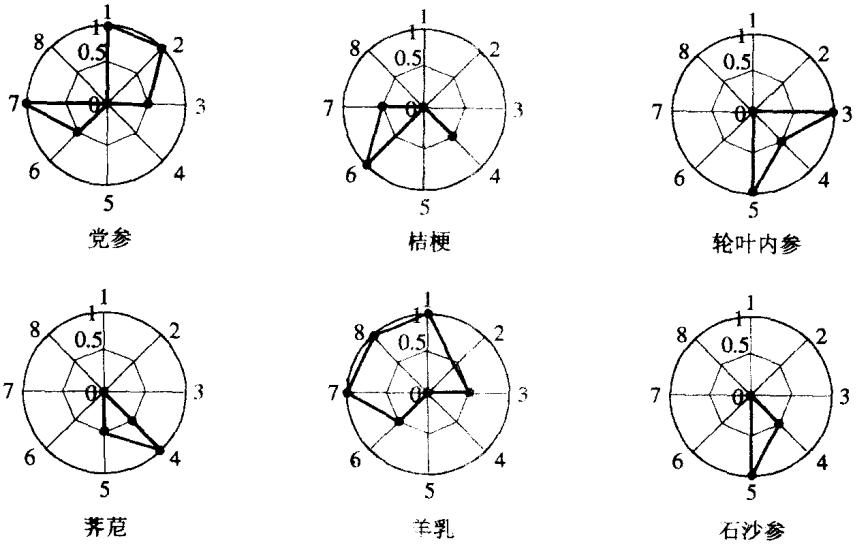

图 1-1 例 1 数据的雷达图

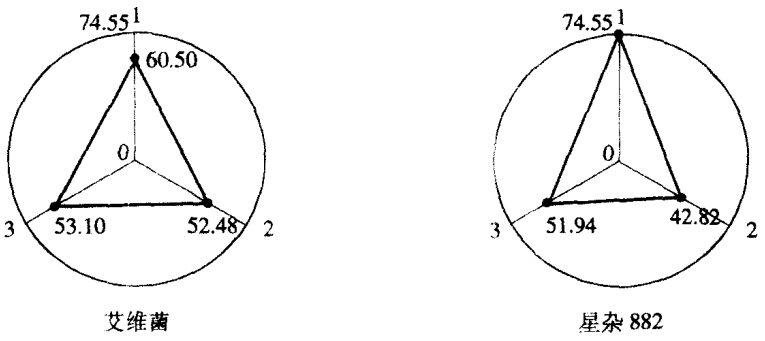

图 1-2 例 2 数据的雷达图

一般来讲，n 个样品 n 个雷达图，如果样品很少，可以画在一个雷达图上，如例 2 的雷达图可以合并. 雷达图可以直观展示各样品的特征及相似差异.

2. 轮廓图

轮廓图是用 p 个平行的纵轴代表 p 个变量，每个样品在图上有 p 个点，将它们依次连接起来成一折线，这个折线图称为样品的轮廓图. 例 1 的轮廓图如图 1-3 所示，例 2 的轮廓图如图 1-4 所示.

除了上述两种图示法之外，其它的多元数据图示法还有塑像图、树形图、星座图、脸谱图和三角多项式图等. 上述这些多元数据图示法，只适合于 p 个变量平等的情形，不能表示出变量间的相关关系. 要想表示出变量间的相关规律，还有其它

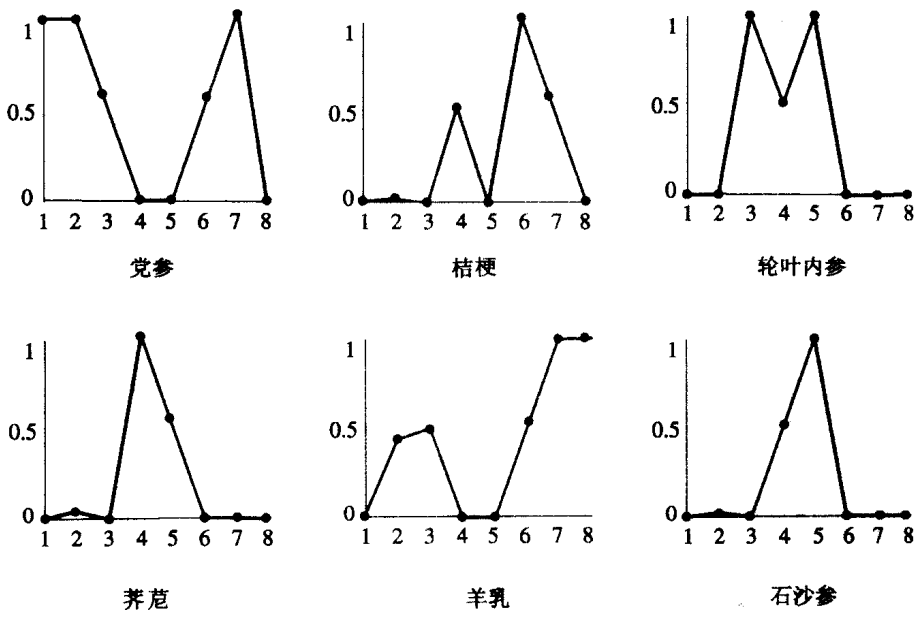

图 1-3　例 1 数据的轮廓图

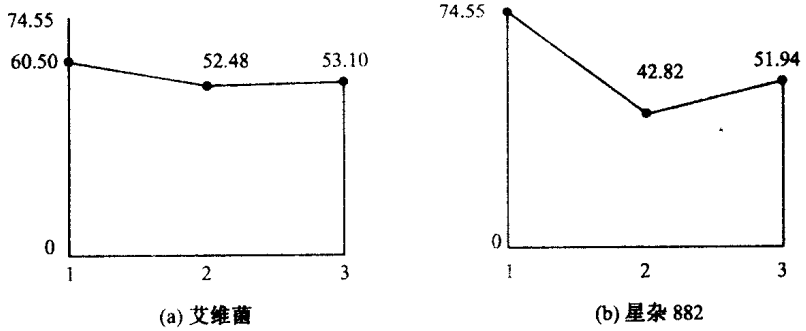

图 1-4　例 2 数据的轮廓图

的图示方法，如连结向量图等，由于这些方法比较复杂，这里就不介绍了.

§1.2　多元正态分布

在农业科学研究中，经常通过多元统计数据寻求多维随机向量的统计规律问题，通俗地讲，就是多指标统计问题. 如，自然因子对作物产量和品质的作用、育种中选择性状对目标性状的影响、疾病的多指标诊断、土壤成分分析、气象预报、样品

的归属等，都属于多指标统计问题.这些指标间千丝万缕、交叉影响，存在着极其复杂的统计规律.在这种情况下，用我们所熟知的一元统计知识去孤立地分析各个指标，就难免顾此失彼，使整体结论失真.因而，多指标、多因素问题只有选择相应的多元统计方法来处理，才能使其规律得以正确的表达.

多元统计分析的主要理论都是建立在多元正态分布总体基础上的.在实际问题中，所遇到的多元总体多是多元正态分布总体或近似多元正态分布，有时不是多元正态分布总体，但当样本容量足够大时，其平均值将近似服从多元正态分布.

1.2.1 多元正态分布的定义

设某品种小麦的产量(X_1)、每亩穗数(X_2)、每穗粒数(X_3)和千粒重(X_4)均服从正态分布：

$$X_i \sim N(\mu_1, \sigma_i^2), \qquad i=1,2,3,4 \tag{1.2.1}$$

其中 μ_i 与 σ_i^2 分别为 X_i 的均值与方差.其密度函数为

$$\begin{aligned} f(x_i) &= \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left\{-\frac{1}{2\sigma_i^2}(x_i-\mu_i)^2\right\} \\ &= (2\pi)^{-\frac{1}{2}}(\sigma_i^2)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(x_i-\mu_i)(\sigma_i^2)^{-1}(x_i-\mu_i)\right\}, \quad -\infty<X_i<+\infty \end{aligned} \tag{1.2.2}$$

那么四个性状所组成的四维列向量 $X=[X_1, X_2, X_3, X_4]^{\mathrm{T}}$ 就服从四元正态分布.

一般地，对于每一个分量都服从正态分布的 p 维随机列向量

$$X=[X_1, X_2, \cdots, X_p]^{\mathrm{T}} \tag{1.2.3}$$

具有和一元正态分布相似的概率密度函数

$$f(X)=(2\pi)^{-p/2}|\Sigma|^{-1/2}\exp\{-1/2(X-\mu)^{\mathrm{T}}\Sigma^{-1}(X-\mu)\} \tag{1.2.4}$$

其中，

$$X=(x_1, x_2, \cdots, x_p)^{\mathrm{T}}, \quad -\infty<x_i<+\infty, \quad i=1,2,\cdots,p$$

$$\mu=(\mu_1, \mu_2, \cdots, \mu_p)^{\mathrm{T}},$$

$$\Sigma=\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix}, \text{其行列式} |\Sigma|>0$$

这时称 X 服从 p 元正态分布，记作 $X \sim N_p(\mu, \Sigma)$.

在式(1.2.4)中，μ 为 X 的数学期望向量，即均值向量.E 为期望算子，有 $E(X)=\mu$，μ 的分量 $\mu_i(i=1,2,\cdots,p)$ 为 X 的分量 X_i 的数学期望，即 $E(X_i)=\mu_i$.Σ 称为 X 的协方差阵.V 为方差算子，即 $V(X)=\Sigma$，Σ 中的元素 σ_{ii} 为 X 的分

量 X_i 的方差，即 $V(X_i)=\sigma_i^2=\sigma_{ii}$，而 σ_{ij} 为 X_i 与 X_j 的协方差，$\sigma_{ij}=\sigma_{ji}$. 在(1.2.4)中，要求 $|\Sigma|>0$，即 Σ 不但对称而且正定，这时 Σ 的逆 Σ^{-1} 一定存在，且 Σ 的特征根均大于 0. 正态分布参数的这些性质，在农林科研中都具有特定的实际意义.

和一元正态分布相仿，如果 X 服从 P 元正态分布，而且 $E(X)=0$，$V(X)=I$（I 为 p 阶单位阵），这时称 X 服从 p 元标准正态分布，记为 $X\sim N_p(0,I)$. 其概率密度函数为

$$f(X)=(2\pi)^{-\frac{p}{2}}\exp\left\{-\frac{1}{2}X^{\mathrm{T}}X\right\}. \tag{1.2.5}$$

1.2.2　多元正态分布的性质

多元正态分布有许多优良的性质，在实际应用中起着重要的作用，简述如下：

1. 若 $X\sim N_p(\mu,\Sigma)$，则

$$d^2=(X-\mu)^{\mathrm{T}}\Sigma^{-1}(X-\mu)\sim\chi^2(P) \tag{1.2.6}$$

d^2 若为定值，X 变化，则它为一椭球，是 X 密度函数的等高面. 若 X 给定，则 d^2 为 X 到 μ 的马哈拉诺比斯距离.

2. 若 $X\sim N_p(\mu,\Sigma)$，则它任一 $K(K\leqslant p)$ 维子向量服从 K 维正态分布，其均值向量由子向量各分量的均值组成，其协方差阵由子向量各分量的方差及协方差组成.

由性质 2 知，X 的任一分量 $X_i\sim N(\mu_i,\sigma_i^2)$.

3. 若 $X\sim N_p(\mu,\Sigma)$，若将 X 分割为两个列向量 Y_1 和 Y_2，即 $X=(Y_1,Y_2)^T$，其均值向量亦分割为相应的两个列向量 $\mu=(\mu_1,\mu_2)^{\mathrm{T}}$. Y_1，Y_2 的协方差阵分别为 Σ_{11}，Σ_{22}. Y_1 与 Y_2 间的协方差阵 $\mathrm{cov}(Y_1,Y_2)=\Sigma_{12}$，而 $\mathrm{cov}(Y_2,Y_1)=\Sigma_{21}$，$\Sigma_{12}{}^{\mathrm{T}}=\Sigma_{21}$. 则 Y_1 与 Y_2 相互独立的充要条件为 $\Sigma_{12}=0$.

在生物科学中，X 的各个分量之间存在着复杂的相关关系. 有时 X 可以分为若干个低维子向量，它们各自代表着不同的功能性状组. 性质 3 为我们提出了解决此类问题的方向.

4. 若 $X\sim N_p(\mu,\Sigma)$，C 为 $q\times p$ 阶非零常数矩阵，b 为 q 维常数列向量，则

$$Y=CX+b\sim N_g(C\mu+b,C\Sigma C^{\mathrm{T}}) \tag{1.2.7}$$

它表明多元正态分布经过线性变换仍为正态分布. 特别地，当 $q=1$，$b=0$ 时，Y 为 X 各分量的线性组合，它服从一元正态分布，即

$$Y=\sum_{i=1}^{p}C_iX_i\sim N\Big(\sum_{i=1}^{p}C_i\mu_i,\sum_{i=1}^{p}\sum_{j=1}^{p}C_iC_j\sigma_{ij}\Big)$$

这为我们研究多个变量与多个变量之间的关系带来了极大的方便.

5. 若 $X\sim N_p(\mu,\Sigma)$，对 X 的各个分量作标准变换

$$Y_i = \frac{X_i - \mu_i}{\sigma_i}, \qquad i = 1, 2, \cdots, p$$

则 $Y = [Y_1, Y_2, \cdots, Y_p]^{\mathrm{T}} \sim N_p(0, \rho)$，即服从 P 元正态分布，其中 0 为 P 维零列向量，ρ 为 X 的相关阵：

$$\rho = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix},$$

其中 $\rho_{ij} = \sigma_{ij}/\sigma_i\sigma_j$ 为 X_i 与 X_j 的相关系数.

§1.3 多元正态分布参数的估计

由样本推断总体，是统计学的基本特点与方法. 推断的主要内容之一就是总体参数的估计.

1.3.1 样本

设从总体 $X \sim N_p(\mu, \Sigma)$ 中随机抽取容量为 n 的多元随机样本 $X_{(j)} = (X_{1j}, X_{2j}, \cdots, X_{pj})^{\mathrm{T}}$，$j = 1, 2, \cdots, n$，$n > p$，由于是随机抽样，故 $X_{(j)}$ 之间相互独立且均服从 $N_p(\mu, \Sigma)$. 样本用矩阵表示为

$$X = (X_{(1)}, X_{(2)}, \cdots, X_{(n)}) = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & & \vdots \\ X_{p1} & X_{p2} & \cdots & X_{pn} \end{bmatrix} \tag{1.3.1}$$

称为观察矩阵或样本资料阵. 在理论上观察阵 X 为随机矩阵，第 j 列 $X_{(j)}$ 为第 j 个样品 p 个指标的 p 维列向量；当测定后为一数据阵，此时将 X_{ij} 换为 x_{ij}. 观察阵中包含了样本对于总体的所有信息，统计分析就要从中科学地提取这些信息，达到认识总体的目的.

1.3.2 样本的数字特征

由样本观察矩阵(1.3.1)，通过计算可得样本的数字特征.

1. 样本均值向量

样本均值向量 $\overline{X}$ 是表示样本中心位置的，其定义为

$$\overline{X} \triangleq \frac{1}{n}\sum_{j=1}^{n} X_{(j)} = \frac{1}{n}\begin{bmatrix} X_{11}+X_{12}+\cdots+X_{1n} \\ X_{21}+X_{22}+\cdots+X_{2n} \\ \vdots \\ X_{p1}+X_{p2}+\cdots+X_{pn} \end{bmatrix} = \begin{bmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \vdots \\ \overline{X}_p \end{bmatrix} \tag{1.3.2}$$

2.样本离差阵

样本离差阵 L 亦称为样本信息阵,因为它反映了样本中各指标及指示间的变异及相关信息,令

$$\widetilde{X}_{(j)} = X_{(j)} - \overline{X}$$

$$\widetilde{X} = (\widetilde{X}_{(1)}, \widetilde{X}_{(2)}, \cdots, \widetilde{X}_{(n)}) = \begin{bmatrix} X_{11}-\overline{X}_1 & X_{12}-\overline{X}_1 & \cdots & X_{1n}-\overline{X}_1 \\ X_{21}-\overline{X}_2 & X_{22}-\overline{X}_2 & \cdots & X_{2n}-\overline{X}_2 \\ \vdots & \vdots & & \vdots \\ X_{p1}-\overline{X}_p & X_{p2}-\overline{X}_p & \cdots & X_{pn}-\overline{X}_p \end{bmatrix}$$

样本的离差阵定义为

$$L \triangleq \widetilde{X}\widetilde{X}^{\mathrm{T}} = \sum_{j=1}^{n}(X_{(j)}-\overline{X})(X_{(j)}-\overline{X})^{\mathrm{T}}$$

$$= \sum_{j=1}^{n} X_{(j)}X_{(j)}^{\mathrm{T}} - \frac{1}{n}\Big(\sum_{j=1}^{n} X_{(j)}\Big)\Big(\sum_{j=1}^{n} X_{(j)}\Big)^{\mathrm{T}} = \sum_{j=1}^{n} X_{(j)}X_{(j)}^{\mathrm{T}} - n\overline{X}\,\overline{X}^{\mathrm{T}} \tag{1.3.3}$$

令

$$L = \begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1p} \\ l_{21} & l_{22} & \cdots & l_{2p} \\ \vdots & \vdots & & \vdots \\ l_{p1} & l_{p2} & \cdots & l_{pp} \end{bmatrix}$$

其中

$$\begin{cases} l_{ii} = \sum_{j=1}^{n}(X_{ij}-\overline{X}_i)^2 = \sum_{j=1}^{n} X_{ij}^2 - \frac{1}{n}\Big(\sum_{j=1}^{n} X_{ij}\Big)^2 \\ l_{ij} = \sum_{k=1}^{n}(X_{ik}-\overline{X}_i)(X_{jk}-\overline{X}_j) = \sum_{k=1}^{n} X_{ik}X_{jk} - \frac{1}{n}\Big(\sum_{k=1}^{n} X_{ik}\Big)\Big(\sum_{k=1}^{n} X_{jk}\Big) \end{cases} \tag{1.3.4}$$

l_{ii} 称为第 i 个分量 X_i 的样本偏差平方和,l_{ij} 称为 X_i 与 X_j 的偏差积和.

样本的均值向量和离差阵可直接根据观察阵计算,令 $I_n = (1,1,\cdots,1)^{\mathrm{T}}$,$I_n$ 为 n 阶单位阵,则有

$$\begin{cases} \overline{X} = \frac{1}{n}XI_n \\ L = X(I_n - I_nI_n^{\mathrm{T}})X^{\mathrm{T}} \end{cases} \tag{1.3.5}$$